**IN THE CLAIMS** – Following is the list of claims and their status:

1. (Currently Amended) A computer-assisted method for identifying duplicate and near-duplicate documents in a large collection of documents, comprising the steps of:

initially, selecting distinctive features contained in the collection of documents,

then, for each document, identifying the distinctive features contained in the document, and

then, for each pair of documents having at least one distinctive feature in common, comparing the distinctive features of the documents to determine whether the documents are duplicate or near-duplicate documents,

wherein the distinctive features are text fragments, which are sequences of at least two words that appear in a limited number of documents in the document collection,

wherein the text fragments are determined to be distinctive features based upon a function of the frequency of a text fragment within a document in the large collection of documents,

wherein for each sequence of at least two words, a distinctiveness score is calculated, and the highest scoring sequences that are found in at least two documents in the document collection are considered distinctive text fragments,

wherein the distinctiveness score is the reciprocal of the number of documents containing the text fragment multiplied by a monotonic function of the number of words in the text fragment.

2. (Currently Amended) The computer-assisted method according to claim 1, wherein the method is applied to: removing duplicates in document collections; detecting plagiarism, detecting copyright infringement; determining the authorship of a document; clustering successive versions of a document from among a collection of documents; seeding a text classification or text clustering algorithm with sets of duplicate or near-duplicate documents; matching an e-mail message with responses to the e-mail message; matching responses to an e-mail message with the e-mail message; creating a document index for use with a query system to efficiently find documents in response to a query which contain a particular phrase or excerpt, or any combination thereof.

3-10. (Cancelled)

11. (Currently Amended) The computer-assisted method according to claim [[~~10~~]] 2, wherein the document index can be utilized even if the particular phrase or excerpt was not recorded correctly in the document or in the query.

12. (Original) The computer-assisted method according to claim to 1, wherein the distinctive features appear in a different order in each of the documents.

13. (Cancelled)

14. (Currently Amended) The computer-assisted method according to claim

1 [[~~13~~]], wherein the method is applied to information retrieval methods.

15. (Previously Presented) The computer-assisted method according to claim 14, wherein a text classification method is applied to the information retrieval method.

16. (Original) The computer-assisted method according to claim 14, wherein:

the information retrieval method assumes word independence, and

the distinctive text fragments are added to an index set.

17. (Cancelled)

18. (Previously Presented) The computer-assisted method according to claim 1, wherein if one distinctive text fragment is contained within another distinctive text fragment within the same document, only the longest distinctive text fragment is considered as a distinctive feature.

19. (Cancelled)

20. (Previously Presented) The computer-assisted method according to claim 1, wherein the sequences of at least two words are considered as appearing in a document when the document contains the sequence of a user-specified minimum frequency.

21.-22. (Cancelled)

23. (Currently Amended) The computer-assisted method according to claim [[~~22~~]] 1, wherein the monotonic function is the number of words in the text fragment.

24. (Cancelled)

25. (Currently Amended) The computer-assisted method according to claim [[~~24~~]] 40, wherein the monotonic function is the number of words in the text fragment.

26. (Previously Presented) The computer-assisted method according to claim 1, wherein the limited number is selected by a user.

27. (Previously Presented) The computer-assisted method according to claim 1, wherein the limited number is defined by a linear function of the number of documents in the document collection.

28. (Previously Presented) The computer-assisted method according to claim 1, wherein the distinctive text fragments include glue words.

29. (Original) The computer-assisted method according to claim 28, wherein

the glue words do not appear at either extreme of the distinctive text fragments.

30. (Currently Amended) ~~The~~ A computer-assisted method ~~according to claim 1~~ for identifying duplicate and near-duplicate documents in a large collection of documents, comprising the steps of:

initially, selecting distinctive features contained in the collection of documents,

then, for each document, identifying the distinctive features contained in the document, and

then, for each pair of documents having at least one distinctive feature in common, comparing the distinctive features of the documents to determine whether the documents are duplicate or near-duplicate documents,

wherein the distinctive features are text fragments, which are sequences of at least two words that appear in a limited number of documents in the document collection,

wherein the text fragments are determined to be distinctive features based upon a function of the frequency of a text fragment within a document in the large collection of documents,

further including the step of, for each pair of documents having at least one distinctive feature in common, counting the number of distinctive features in common,

wherein determining whether the pair of documents is duplicates or near-duplicates includes the steps of:

for each pair of documents, calculating an overlap ratio by dividing the number of distinctive features in common by the smaller of the number of distinctive features per document, and

comparing the overlap ratio to a threshold and if the overlap ratio is greater than the threshold, then the pair of documents are duplicates or near-duplicates, otherwise the pair of documents is not duplicates or near-duplicates,

building a document index that maps each document to its associated distinctive features, wherein if one distinctive feature is repeated within one document, the index maps the document to the distinctive feature once, and

building a feature index that maps each distinctive feature to its associated document, wherein if one distinctive feature is repeated within one document, the index maps the distinctive feature to the document once,

wherein determining whether the pair of documents are duplicates or near-duplicates further includes the steps of:

creating a list of unique distinctive features from the document index,

for each unique distinctive feature, creating a list of documents which contain the unique distinctive feature, and

for each document, creating a list of documents that have at least one feature in common with the document and the number of features in common with the document.

31. (Cancelled)

32. (Currently Amended) The computer-assisted method according to claim [[~~31~~]] 30, wherein the distinctive features include distinctive phrases.

33. (Currently Amended) The computer-assisted method according to claim [[~~31~~]] 30, wherein the distinctive features appear in a different order in each of the documents.

34. (Currently Amended) The computer-assisted method according to claim [[~~31~~]] 30, wherein the distinctive features include text spans.

35. (Original) The computer-assisted method according to claim 34, wherein the text spans include sentences.

36. (Original) The computer-assisted method according to claim 34, wherein the text spans include lines of text.

37.-39. (Cancelled)

40. (NEW) A computer-assisted method for identifying duplicate and near-duplicate documents in a large collection of documents, comprising the steps of:

initially, selecting distinctive features contained in the collection of documents,

then, for each document, identifying the distinctive features contained in the document, and

then, for each pair of documents having at least one distinctive feature in common, comparing the distinctive features of the documents to determine whether the documents are duplicate or near-duplicate documents,

wherein the distinctive features are text fragments, which are sequences of at least two words that appear in a limited number of documents in the document collection,

wherein the text fragments are determined to be distinctive features based upon a function of the frequency of a text fragment within a document in the large collection of documents,

wherein for each sequence of at least two words, a distinctiveness score is calculated, and the highest scoring sequences that are found in at least two documents in the document collection are considered distinctive text fragments,

wherein the distinctiveness score is the percentage of documents not containing the phrase multiplied by a monotonic function of the number of words in the  text fragment.

41. (NEW) The computer-assisted method according to claim 40, wherein the method is applied to: removing duplicates in document collections; detecting plagiarism, detecting copyright infringement; determining the authorship of a document; clustering successive versions of a document from among a collection of documents; seeding a text classification or text clustering algorithm with sets of duplicate or near-duplicate documents; matching an e-mail message with responses to the e-mail message; matching responses to an e-mail message with the e-mail message; creating a document index for use with a query system to efficiently find documents in response to a query which contain a particular phrase or excerpt, or any combination thereof.

42. (NEW) The computer-assisted method according to claim 41, wherein the document index can be utilized even if the particular phrase or excerpt was not recorded correctly in the document or in the query.

43. (NEW) The computer-assisted method according to claim to 40, wherein the distinctive features appear in a different order in each of the documents.

44. (NEW) The computer-assisted method according to claim 40, wherein the method is applied to information retrieval methods.

45. (NEW) The computer-assisted method according to claim 44, wherein a text classification method is applied to the information retrieval method.

46. (NEW) The computer-assisted method according to claim 44, wherein:

the information retrieval method assumes word independence, and

the distinctive text fragments are added to an index set.

47. (NEW) The computer-assisted method according to claim 40, wherein if one distinctive text fragment is contained within another distinctive text fragment within the same document, only the longest distinctive text fragment is considered as a distinctive feature.

48. (NEW) The computer-assisted method according to claim 40, wherein the

sequences of at least two words are considered as appearing in a document when the document contains the sequence of a user-specified minimum frequency.

49. (NEW) The computer-assisted method according to claim 40, wherein the limited number is selected by a user.

50. (NEW) The computer-assisted method according to claim 40, wherein the limited number is defined by a linear function of the number of documents in the document collection.

51. (NEW) The computer-assisted method according to claim 40, wherein the distinctive text fragments include glue words.

52. (NEW) The computer-assisted method according to claim 51, wherein the glue words do not appear at either extreme of the distinctive text fragments.